Research Article

Psychological Science
1–14

DOI: 10.1177/0956797618780658
www.psychologicalscience.org/PS

The Paradox of Viral Outrage

Takuya Sawaoka[1] and Benoît Monin[1,2]
[1]Department of Psychology and [2]Graduate School of Business, Stanford University

Abstract

Moral outrage has traditionally served a valuable social function, expressing group values and inhibiting deviant behavior, but the exponential dynamics of Internet postings make this expression of legitimate individual outrage appear excessive and unjust. The same individual outrage that would be praised in isolation is more likely to be viewed as bullying when echoed online by a multitude of similar responses, as it then seems to contribute to disproportionate group condemnation. Participants (N = 3,377) saw racist, sexist, or unpatriotic posts with accompanying expressions of outrage and formed impressions of a single commenter. The same commenter was viewed more negatively when accompanied by a greater number of commenters (i.e., when outrage was viral vs. nonviral), and this was because viral outrage elicited greater sympathy toward the initial offender. We examined this effect and its underlying processes across six studies.

Keywords

morality, outrage, social judgment, social justice, punishment, open data, open materials, preregistered

Received 9/24/17; Revision accepted 5/8/18

Note: This article reproduces offensive images and language that were used as stimuli in the studies reported.

Moral outrage—a combination of anger and disgust at the violation of a moral standard—is increasingly common in contemporary public discourse (Batson & Shaw, 1991; Brady, Wills, Jost, Tucker, & Van Bavel, 2017; Crockett, 2017; Haidt, 2003; Hechler & Kessler, 2018; Rozin, Lowery, Imada, & Haidt, 1999; Salerno & Peter-Hagene, 2013). Online remarks deemed offensive often trigger an outpouring of outrage extending far beyond the poster's immediate social circle. Observers typically admire people who take a stance against injustice (Walker & Hennig, 2004), who refuse to complete a racist task (Monin, Sawyer, & Marquez, 2008), or who punish free riders in public-goods games (Barclay, 2006; Raihani & Bshary, 2015). Confronting injustice, from this perspective, is a praiseworthy and righteous act (Rattan & Dweck, 2010) that sustains human cooperation (Czopp, Monteith, & Mark, 2006; Fehr & Gächter, 2002).

However, we suggest that this is not the whole story; outrage can be perceived as going too far when such perfectly justified condemnation accumulates into a phenomenon referred to as *viral outrage* (e.g., Hempel, 2017). In one famous example, after Justine Sacco tweeted, "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!" her post unexpectedly went viral, eliciting rebukes from thousands ("You racist"; "I am beyond horrified"; "What the hell is wrong with you?"). This widespread online shaming severely damaged Sacco's personal and professional life. Others later expressed sympathy toward Sacco, arguing that although her tweet was offensive, her detractors took the outrage too far. These observers believed that while many of those who responded to Sacco may have earnestly been trying to combat racism, the cumulative consequence of their actions on Sacco's life made them bullies, not heroes (Ronson, 2015).

In this article, we document and seek to explain these puzzling reactions to moral outrage. At first glance, the phenomenon we describe seems at odds

Corresponding Author:
Takuya Sawaoka, Stanford University, Department of Psychology, 450 Serra Mall, Jordan Hall, Stanford, CA 94305
E-mail: sawaoka@stanford.edu

with theories of conformity (Asch, 1956) and demonstrations of the power of social proof (Cialdini, 1993): If 10 people accuse a speaker of racism, each accuser should be perceived as more legitimate than if only 2 do so. But we propose that when people are deciding whether outrage and condemnation are appropriate or uncalled for, they also consider how the target of outrage feels in response to the condemnation. If 10 people say that a line is 12 in. long, they each seem justified in their judgment, but when 10 people denounce someone as racist, observers may start to feel sympathy for the offender. The final outcome is that when evaluating any individual expression of outrage, people may view it as less proportionate to the offense when they see it as part of an unjust whole. This is the paradox of viral outrage: The exact same individual expression of outrage may appear laudable in isolation but morally suspect when accompanied by a chorus of echoing outrage.

We documented this effect and cast light on the underlying processes in six studies. First, we demonstrated that the same expression of outrage against a racist (Studies 1, 4b, and 5), unpatriotic (Study 2), or sexist (Studies 3 and 4a) online post is seen as appropriate in isolation but less praiseworthy in the context of many others echoing this condemnation. Second, we showed that this mistrust of viral outrage results from observers feeling increasingly sympathetic for the original offender as the outrage grows. We did so by measuring sympathy as a mediator (Studies 2–5) and by manipulating it directly (Studies 4a and 4b). Third, we ruled out multiple alternative mechanisms for these effects (Studies 2 and 3). Fourth, we demonstrated that individuals expressing outrage were not affected in their perception of the legitimacy of their own outrage by the number of others accompanying them (Study 5), suggesting a potential disconnect with the judgments of potential observers.

General Method

Target sample sizes were determined a priori for all studies. This target was set at 100 per cell (before participant exclusions) for Studies 1, 2, 3, and 5, exceeding the target sample size recommended by Simmons, Nelson, and Simonsohn (2013). Studies 4a and 4b were conducted during the revision process to address reviewer comments; to avoid ambiguous results at that stage, we preregistered them at https://osf.io/kges3/ and increased our sample size a priori to 200 per cell. Across all studies, we always excluded a priori participants whose Internet protocol (IP) address appeared more than once. All of our data are publicly available on the Open Science Framework at https://osf.io/qtsx2/. Across the main text and the Supplemental Material available online, we report all measures, conditions, and data exclusions for all studies. To help readers better evaluate the reliability of our findings, we have compiled a list of the *p* values and effect sizes obtained in each study (Table 1).

Table 1. Effect Sizes and *p* Values for the Effect of Viral (vs. Nonviral) Outrage on Impressions of the Target Commenter

Study	*p*	Effect size (*d*)
Study 1	.004	0.29
Study 2	.001	0.29
Study 3	.011	0.21
Study 4a	< .001	0.25
Study 4b	.046	0.14
Study 5	.008	0.35

Note: The values for Study 5 refer to the effect of viral (vs. nonviral) outrage in the third-person-observer condition. An additional study ($N = 810$) was included in the original manuscript but was removed in the revision process to avoid redundancy with the present Studies 4a and 4b. Results of this omitted study are in line with the results presented here ($p = .006$, $d = 0.20$ for the focal effect of viral outrage on negative impressions).

In our studies, we first presented participants with an offensive post taken directly from or inspired by actual posts that led to viral outrage. We operationalized viral (vs. nonviral) outrage by manipulating the number of outraged comments reacting to these posts. In the nonviral-outrage condition, participants read 2 responses, whereas in the viral-outrage condition they read 10. Although viral posts often receive thousands of responses, minimal manipulations of independent variables offer more stringent tests of hypothesized effects (Prentice & Miller, 1992), demonstrating how simply and subtly an effect can be produced. We also reasoned that any observed differences between 2 and 10 responses would also emerge between 2 and 10,000 responses, for example. After reading these outraged responses, participants formed impressions of a specific individual who took part in the outrage.

Study 1

In Study 1, participants were presented with an offensive social media post taken from a real news story and then saw a series of outraged responses. Participants received a minimal manipulation of viral (vs. nonviral) outrage and evaluated the moral character of a single commenter. We predicted that participants would rate the target commenter more negatively when there were more commenters—that is, when outrage went viral. We also manipulated whether participants evaluated the first or last commenter, to examine whether the position of the target commenter influenced participants' impressions. We reasoned that while it might make sense to

penalize a latecomer who piled on the recrimination by jumping on the bandwagon, there would be less normative justification for letting the number of comments affect reactions to the first comment posted.

Method

Participants. Participants (N = 397; 147 women; 100 racial-minority individuals; mean age = 32.39 years, SD = 10.09) were recruited and compensated through Amazon's Mechanical Turk. Participants were excluded if their IP address was recorded more than once (n = 7), leaving a final sample of 390.

Procedure. Participants viewed an actual social media post (Fig. 1, top) in which a White woman pictured herself with black tape on her face and joked about fitting into her historically Black college (Clemons, 2016). We did not use the student's real name. Participants indicated how offensive they found this post (1, *not at all*, to 7, *very*). They then read fictitious, condemnatory responses (Fig. 1, bottom) to this post inspired by real responses at the time of the incident. Participants read 2 or 10 responses, depending on whether they were assigned to the nonviral- or viral-outrage conditions, respectively. In both conditions, we specified that these were all of the responses that the post received. Participants were informed that responses were presented in chronological order, and they evaluated either the first or last commenter. We randomized the order in which these responses were presented, as well as the target commenter that participants evaluated. Participants evaluated the target commenter as "in the wrong," "a bully," "praiseworthy," and "a good person" (the latter two were reverse-coded; 1, *not at all*, to 7, *totally*; α = .86).

Results

Participants found the original transgression to be moderately offensive (M = 5.16, SD = 1.86). This did not differ by condition, ps > .191. We conducted a 2 (outrage: nonviral or viral) × 2 (target commenter: first or last) analysis of variance (ANOVA) on impressions of the target commenter. We included the target stimulus (i.e., which of 10 possible outrage stimuli participants evaluated) as a fixed effect. This revealed a main effect of outrage, $F(1, 377) = 8.51$, $p = .004$, $\eta_p^2 = .02$ (Fig. 2); participants formed more negative impressions of the single commenter when outrage was viral (M = 3.66, SD = 1.56) rather than nonviral (M = 3.28, SD = 1.28). There was no effect of whether participants evaluated the first or last commenter, and no interaction, ps > .235. As predicted, an individual's outrage was viewed less favorably when it was echoed by other commenters than when it was presented in relative isolation, even though the other commenters acted independently of the target. These results were unaffected by whether

Fig. 1. The offensive stimulus (top) and examples of outrage stimuli (bottom) used in Studies 1 and 5. The specific comments shown at the bottom were sampled randomly for each participant, with each comment always matched to the same name and profile picture.

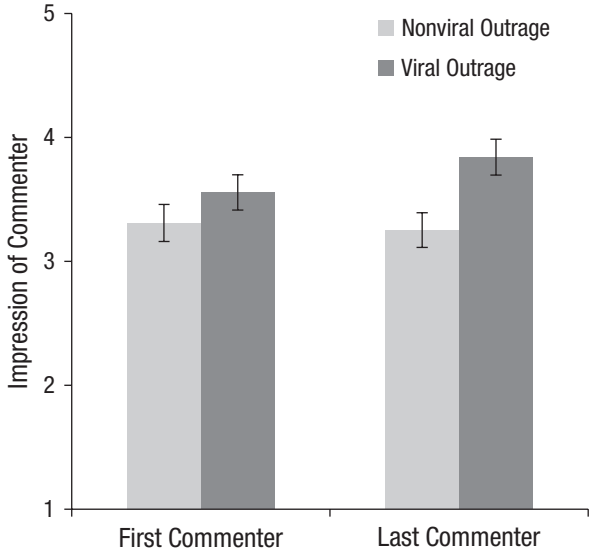

Fig. 2. Results from Study 1: mean impression of target commenter as a function of whether the commenter was first or last and whether the response was viral or nonviral. Impression ratings ranged from 1 to 7; higher scores indicate more negative impressions. Error bars represent standard errors.

participants evaluated the first or last commenter. This is particularly striking, as first commenters presumably had little way of knowing what subsequent commenters would say but were nonetheless penalized for the actions of those who followed in their footsteps.

Study 2

What explains these findings? We propose that targets of viral outrage who are battered with widespread remarks of condemnation and anger elicit sympathy from observers. This sympathy taints perceptions of individual commenters, making them look less like heroes and more like bullies. In Study 2, we tested for the mechanism behind this reaction by measuring sympathy toward the offender as a mediator.

One alternative explanation is that people who participate in viral outrage are perceived to be grandstanding; that is, jumping on the outrage bandwagon to receive reputational benefits in the eyes of others (Tosi & Warmke, 2016). Such cynical attributions for others' apparently moral behavior can lead to backlash (Inesi, Gruenfeld, & Galinsky, 2012). This alternative is weakened by the fact that we did not obtain order effects in Study 1, but we addressed this alternative more directly in Study 2. We included a condition in which comments were anonymous (preventing the inference of grandstanding) and used only targets who were the first commenters (and thus could not be following others). We also included a condition in which commenters could support the initial outrage without adding to it with their own comment.

Method

Participants. Participants (N = 602; 314 women; 149 racial-minority individuals; mean age = 35.01 years, SD = 11.88) were recruited and compensated through Mechanical Turk. Participants were excluded if their IP address was recorded more than once (n = 15), leaving a final sample of 587.

Procedure. Participants saw a social media post (Fig. 3, top) taken from a real story. A charity worker posted a photograph on Facebook in which she pretended to shout and make an obscene gesture next to a sign that read "Silence and Respect" at Arlington National Cemetery (Ronson, 2015). We did not use this individual's real name and instead used the fictional name "Sharon." Participants indicated how offensive they found this post and then received the minimal manipulation of viral (vs. nonviral) outrage in which they read responses from either 10 or 2 other users, respectively (Fig. 3, bottom). In a control condition, these responses included commenters' names, as in Study 1. In a new, anonymous condition, participants were advised as follows: "On this social media platform, users have the choice to indicate their names or to remain anonymous." All responses in this condition were anonymous. In a third, upvoting condition, participants were advised as follows:

> On this social media platform, users are able to indicate their support for posts that they agree with or appreciate (similar to "likes" on Facebook or "upvotes" on Reddit). The number next to the red arrow indicates the number of users who supported this particular message.

Participants in this condition saw only one condemnatory response that was upvoted by either one or nine additional users. Two items (averaged together; r = .91) assessed our proposed mediator: "To what extent did you feel sorry for Sharon?" and "To what extent did you feel bad for Sharon?" (1, *not at all*, to 7, *very*). Then, participants evaluated the first commenter using the same items as in Study 1 (α = .87). We randomized the order in which these responses were presented, as well as the target commenter that participants evaluated (i.e., which commenter was assigned to be in the first position).

Results

Participants found the original transgression to be moderately offensive (M = 5.42, SD = 1.98; 1, *not at all*, to 7,

Nonviral-Outrage Condition

Alexander Clark
This is offensive and incredibly tasteless.

Jack Lee
You're going to hell for this.

Anne Burke
You are just pure evil.

Nancy Canzonieri
Absolutely disgusting. Shame on you.

Timothy Bailey
This is completely unacceptable.

Jennifer Coyle
You are a selfish and entitled brat, and deserve to be punished.

Lindsey Fowler
Inexcusable. People who disrespect veterans like this make me sick.

Robert Saunders
This kind of disrespect cannot go unpunished.

Marla Fuentes
She is a disgrace to America.

Dan Boswell
Send this idiot to prison.

Viral-Outrage Condition

Fig. 3. The offensive stimulus (top) and examples of outrage stimuli (bottom) used in Study 2. The specific comments shown at the bottom were sampled randomly for each participant, with each comment always matched to the same name and profile picture.

very). Unexpectedly, participants found the transgression somewhat less offensive in the viral-outrage condition ($M = 5.27$, $SD = 2.05$) than in the nonviral-outrage condition ($M = 5.57$, $SD = 1.90$), $F(1, 581) = 3.69$, $p = .055$, $\eta_p^2 = .01$. Because we measured offensiveness prior to any experimental manipulation, this difference by condition indicates a failure of randomization. Perceived offensiveness was significantly correlated with both our mediator and dependent measure (rs $\geq .46$), so we controlled for offensiveness across our analyses, although our findings remained qualitatively the same regardless of whether we did so.

A 2 (outrage: nonviral or viral) × 3 (response type: control, anonymous, or upvoting) analysis of covariance (ANCOVA) on sympathy toward the offender revealed a main effect of outrage, $F(1, 580) = 5.60$, $p = .018$, $\eta_p^2 = .01$. Participants reported more sympathy toward the offender when outrage was viral ($M = 2.40$, $SD = 1.56$) than when it was nonviral ($M = 2.01$, $SD = 1.53$). There was also an unanticipated main effect of response type, $F(1, 580) = 3.72$, $p = .025$, $\eta_p^2 = .01$, with participants reporting more sympathy in the control condition ($M = 2.37$, $SD = 1.59$) than in the anonymous condition ($M = 2.13$, $SD = 1.56$) or upvoting condition ($M = 2.09$, $SD = 1.53$). There was no interaction, $p = .638$.

We next conducted a 2 (outrage: nonviral or viral) × 3 (response type: control, anonymous, or upvoting) ANCOVA on impressions of the individual target commenter, with target stimulus as a fixed effect. We found a main effect of outrage, $F(1, 571) = 11.44$, $p = .001$, $\eta_p^2 = .02$ (Fig. 4); participants formed more negative impressions of the target commenter when outrage was viral ($M = 3.66$, $SD = 1.57$) than when it was nonviral ($M = 3.20$, $SD = 1.56$). No other effects were significant, ps $\geq .547$.

We also found support for our proposed mediation (Fig. 5), 95% confidence interval (CI) for the indirect effect = [0.022, 0.208], path *a*: $b = 0.28$, path *b*: $b = 0.40$. This supports our claim that viral outrage elicits more sympathy toward the offender, which then leads people

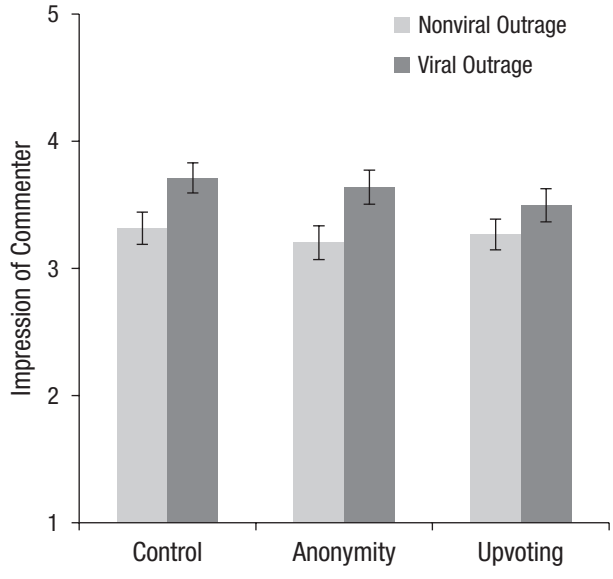

Fig. 4. Results from Study 2: mean impression of target commenter as a function of response type and outrage condition. Impression ratings ranged from 1 to 7; higher scores indicate more negative impressions. Error bars represent standard errors.

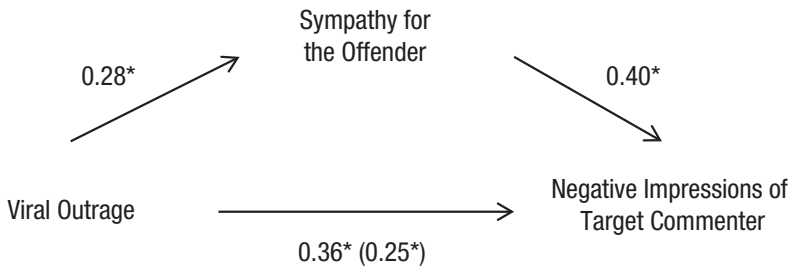

Fig. 5. Model from Study 2 showing the effect of viral outrage on negative impressions of the target commenter, as mediated by sympathy for the offender. Along the bottom path, the value outside parentheses shows the total effect, and the value inside parentheses shows the direct effect after controlling for the mediator. Unstandardized coefficients are shown. Asterisks indicate significant paths ($p < .05$).

to form less favorable impressions of individuals who participate in the outrage. The fact that our findings held even when responses were anonymous rules out the possibility that participants' more negative evaluations arose from the inference that commenters were simply grandstanding.

It is worth noting here that these findings are the opposite of what would be predicted by theories of social proof or third-party punishment. Theories of social proof (Cialdini, 1993) would suggest that 10 (vs. 2) outraged individuals accrue more legitimacy and thus more favorable impressions, but our findings indicate that evaluations of aggregate punishment may diverge fundamentally from perceptions of other aggregated behaviors (e.g., recycling, paying taxes, or not littering; Cialdini, Reno, & Kallgren, 1990). The crucial difference is that in contrast to punishment, these other behaviors would not elicit feelings of sympathy and are unlikely to be perceived negatively as the behaviors become more widespread. Even studies of third-party punishment have typically found positive reputational consequences of those who punish wrongdoers, but to our knowledge, these studies have focused on evaluations of single punishers (e.g., Barclay, 2006), similar to our nonviral-outrage condition. Some researchers have noted that people may not always develop positive perceptions of those who punish (Raihani & Bshary, 2015), consistent with the present findings, and our work highlights how the accumulation of punishment or outrage leads to less favorable impressions of individuals who may have been praised in isolation.

Study 3

A striking feature of our results is that early commenters were de facto penalized for others' independent responses. However, yet another alternative explanation for our findings could be that observers believed that outraged individuals intentionally caused others to follow in their footsteps or, at the very least, that they could foresee that their response would open the floodgates for others' outrage, whether they meant for it to happen or not. This could be the result of observers exaggerating actors' intentional and causal influence (Alicke, 1992; Gray, Young, & Waytz, 2012; Morewedge, 2009). In this view, early expressers of outrage are aware that their comments might elicit many more, and this justifies taking into account later responses when evaluating their own. We addressed this possibility in Study 3. If our findings were driven by observers' assumptions that individual commenters knowingly caused others to follow suit, then our effects should be attenuated when a commenter expresses surprise about the number of other responses—which should signal that this commenter did not expect so many others to join the fray.

Method

Participants. Participants ($N = 604$; 247 women; 144 racial-minority individuals; mean age = 34.86 years, $SD = 16.77$) were recruited and compensated through Mechanical Turk. Participants were excluded if their IP address was recorded more than once ($n = 14$), leaving a final sample of 590.

Procedure. Participants first saw a sexist tweet (Fig. 6, top), adapted from a real story in which a college professor at a conference made comments disparaging female scientists (Quinn, 2015), and indicated how offensive they found this post. They then received the minimal manipulation of viral (vs. nonviral) outrage in which they read 10 or 2 responses, respectively, condemning the original tweet (Fig. 6, bottom). Participants evaluated the first of these responses. Crucially, we manipulated whether target individual commenters expressed surprise at the end about the number of other responses. In the surprise and no-surprise conditions, participants were informed

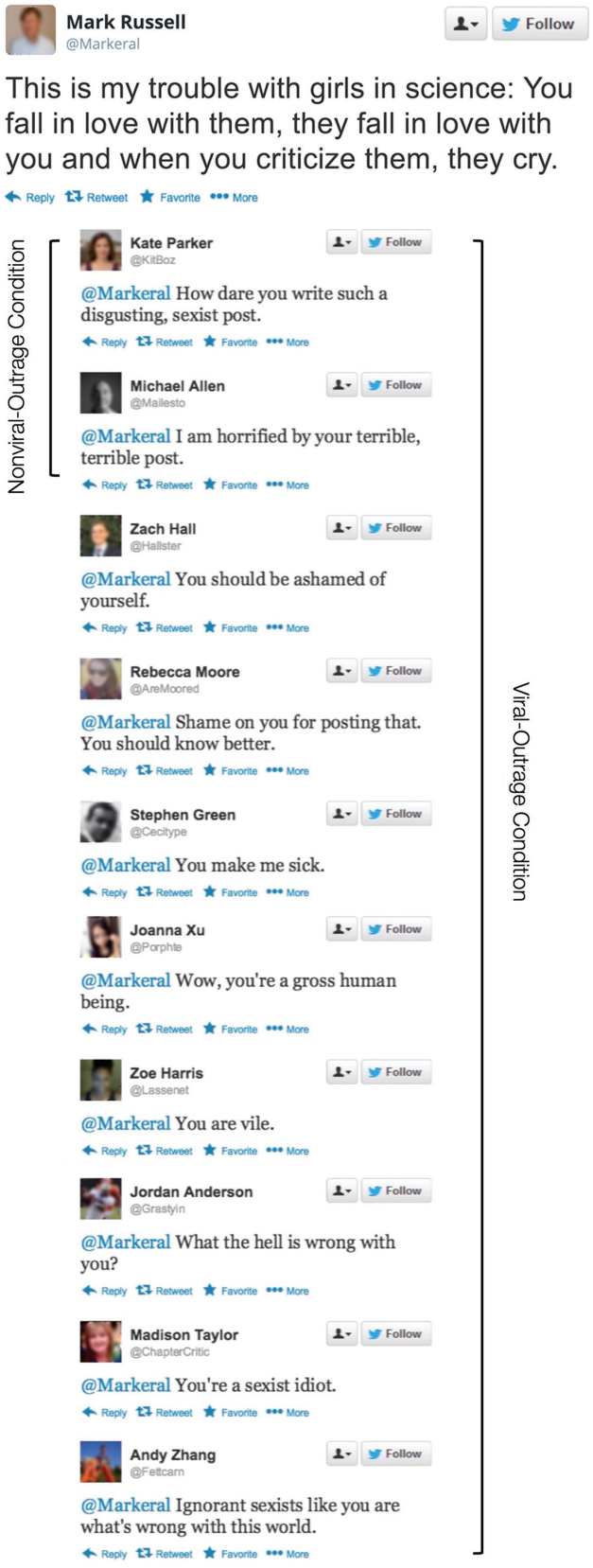

Fig. 6. The offensive stimulus (top) and examples of outrage stimuli (bottom) used in Studies 3 and 4a. The specific comments shown at the bottom were sampled randomly for each participant, with each comment always matched to the same name and profile picture.

that one of the commenters posted a follow-up message. In the surprise condition, this commenter posted either "Wow, I figured I'd be like the only one to respond to this" (viral-outrage condition) or "Wow, I figured like a bunch of people would respond to this" (nonviral-outrage condition). In the no-surprise condition, this commenter posted either "Yeah, I figured like a bunch of people would respond to this" (viral-outrage condition) or "Yeah, I figured I'd be like the only one to respond to this" (nonviral-outrage condition). Participants then indicated their impressions of this target commenter. We again measured sympathy for the offender as a mediator (r = .91). In the control condition, participants did not receive any such instructions, and simply proceeded to indicate their impressions of the target (α = .82).

Results

Participants found the original transgression to be moderately offensive (M = 3.95, SD = 1.96; 1, *not at all*, to 7, *very*). This did not differ by condition, ps ≥ .224. A 2 (outrage: nonviral or viral) × 3 (surprise: control, no surprise, or surprise) ANOVA on sympathy toward the offender revealed a main effect of outrage, $F(1, 584)$ = 5.48, p = .020, η_p^2 = .01; participants felt more sympathy for the offender when outrage was viral (M = 2.71, SD = 1.81) rather than nonviral (M = 2.38, SD = 1.67). There were no effects of surprise, ps ≥ .410.

Next, we conducted a 2 (outrage: nonviral or viral) × 3 (surprise: control, no surprise, or surprise) ANOVA on impressions of the target commenter, including the target stimulus as a fixed effect. This revealed a main effect of outrage, $F(1, 583)$ = 6.57, p = .011, η_p^2 = .01 (Fig. 7), such that participants formed more negative impressions of the target commenter when outrage was viral (M = 3.79, SD = 1.40) rather than nonviral (M = 3.51, SD = 1.34). There was an unexpected main effect of surprise, $F(1, 583)$ = 3.41, p = .034, η_p^2 = .01 (Fig. 7); participants formed more negative impressions of the target commenter in the control condition (M = 3.84, SD = 1.46) than in either the no-surprise condition (M = 3.56, SD = 1.26) or the surprise condition (M = 3.53, SD = 1.39). However, there was no Outrage × Surprise interaction, p = .537, suggesting that our effects were not driven by the perception that target commenters were anticipating the viral outrage that followed. We again found that the effects on impressions of the target commenter were driven by sympathy toward the offender, 95% CI for the indirect effect = [0.023, 0.259], path a: b = 0.33, path b: b = 0.41.

Study 4a

Studies 2 and 3 ruled out several alternative explanations and provided evidence for our proposed mechanism that viral outrage elicits greater sympathy toward

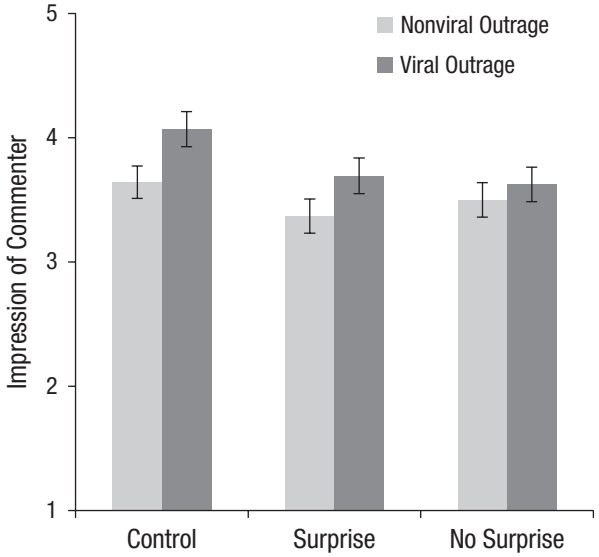

Fig. 7. Results from Study 3: mean impression of target commenter as a function of surprise condition and outrage condition. Impression ratings ranged from 1 to 7; higher scores indicate more negative impressions. Error bars represent standard errors.

the offender, leading to more negative impressions of people who contribute to the aggregate punishment. Next, we further tested this mechanism by manipulating the proposed mediator directly (Spencer, Zanna, & Fong, 2005). If our effects are driven by sympathy, we should be less likely to observe this phenomenon when outrage is directed toward someone that people have difficulty sympathizing with. For example, high-status public figures are typically perceived as more blameworthy for their transgressions (Fragale, Rosen, Xu, & Merideth, 2009) and elicit less sympathy. In Study 4, we manipulated the status of the perpetrator, predicting that the effects of viral outrage would be less likely to emerge for high-status public figures.

Method

Participants. Participants (N = 809; 451 women; 189 racial-minority individuals; mean age = 36.04 years, SD = 11.42) were recruited and compensated through Mechanical Turk. Participants were excluded if their IP address was recorded more than once (n = 10), leaving a final sample of 799. In Study 4a, we set a larger target sample size of 200 per cell, in contrast to our previous studies, because this study was conducted to address questions that arose during the peer-review process. This study was preregistered at https://osf.io/kges3/.

Procedure. Participants saw a tweet taken from a real story (Fig. 8) in which a comedian ridiculed overweight women (Blistein, 2015). We did not use this individual's real name and instead used the fictional name "Michael

Fig. 8. The offensive stimulus used in Study 4a.

Collins." Participants indicated how offensive they found this post. We manipulated the status of the perpetrator by describing him as either an "average Twitter user" (control condition) or an "up-and-coming actor and celebrity who has appeared in several new television programs and a feature film" (high-status-offender condition). Participants then received the minimal manipulation of viral (vs. nonviral) outrage in which they read 10 or 2 responses, respectively, to this individual (the same outrage stimuli used in Study 3). They reported the extent to which they felt sympathetic toward the offender, using the same items as in Studies 2 and 3 (r = .90), and they evaluated a randomly selected commenter using the same items as Studies 1 to 3 (α = .84).

In addition, we included two measures to examine why a high-status offender may elicit less sympathy. Specifically, we measured the extent to which the offender's transgression was perceived as harmful, using four items (e.g., "How much does Michael's tweet contribute to spreading negative perceptions of women?" and "How much does Michael's tweet provide legitimacy to sexist attitudes?"; 1, *not at all*, to 7, *a great deal*; α = .73), and measured the extent to which the offender was perceived to be vulnerable to criticism, using four items (e.g., "How much do you think Michael cares about the comments random people write to him online?" and "How likely do you think Michael is to read the online comments he gets from random people?" 1, *not at all*, to 7, *extremely*; α = .54). We predicted that the offender would be perceived as committing more harm, and as less vulnerable to criticism, when he was of high status (vs. an average Twitter user).

Results

Participants found the original transgression to be moderately offensive (M = 4.90, SD = 1.86; 1, *not at all*, to 7, *very*). Unexpectedly, participants found the transgression more offensive in the viral-outrage condition (M = 5.04, SD = 1.80) than in the nonviral-outrage condition (M = 4.75, SD = 1.92), $F(1, 795) = 4.73$, p = .030, η_p^2 = .01. Because we measured offensiveness prior to any experimental manipulation, this difference by condition indicates a failure of randomization. We thus

controlled for perceived offensiveness across our analyses, although our findings remained qualitatively the same regardless of whether we did so.

First, we tested whether the high-status (vs. control) offender would be perceived as committing more harm and as less vulnerable to criticism. We conducted a 2 (outrage: viral or nonviral) × 2 (offender: high status or control) ANCOVA on perceived harm. This revealed a main effect of offender, $F(1, 794) = 43.29$, $p < .001$, $\eta_p^2 = .05$; participants believed the transgression was more harmful when the offender was of high status ($M = 5.00$, $SD = 1.30$) than in the control condition ($M = 4.52$, $SD = 1.34$). There were no effects of outrage, ps ≥ .121. We then conducted a 2 (outrage: viral or nonviral) × 2 (offender: high status or control) ANCOVA on perceived vulnerability. This revealed a main effect of outrage, $F(1, 794) = 3.89$, $p = .049$, $\eta_p^2 = .01$; participants believed that the offender was more vulnerable when outrage was viral ($M = 3.23$, $SD = 0.99$) rather than nonviral ($M = 3.09$, $SD = 0.99$). Contrary to predictions, there were no effects of offender, ps ≥ .420.

Next, we turned to our main prediction that viral outrage would increase sympathy toward the offender in the control condition but not when the offender was of high status. We conducted a 2 (outrage: viral or nonviral) × 2 (offender: high status or control) ANCOVA on sympathy toward the offender. This revealed a main effect of outrage, $F(1, 794) = 16.85$, $p < .001$, $\eta_p^2 = .02$; participants felt more sympathetic when outrage was viral ($M = 1.93$, $SD = 1.40$) rather than nonviral ($M = 1.62$, $SD = 1.21$). Contrary to predictions, there was no main effect of offender and no interaction, ps ≥ .917.

Finally, we conducted a 2 (outrage: viral or nonviral) × 2 (offender: high status or control) ANCOVA on impressions of the target commenter, with target stimulus as a fixed factor. This revealed a main effect of outrage, $F(1, 785) = 12.29$, $p < .001$, $\eta_p^2 = .02$ (Fig. 9); participants reported more negative impressions of the target commenter when outrage was viral ($M = 3.17$, $SD = 1.47$) rather than nonviral ($M = 2.97$, $SD = 1.30$). Again contrary to predictions, there was no main effect of offender and no interaction, ps ≥ .874. The effect of viral (vs. nonviral) outrage on negative impressions of the target commenter were mediated by sympathy toward the offender, 95% CI for the indirect effect = [0.088, 0.253], path *a*: $b = 0.37$, path *b*: $b = 0.45$.

These findings replicate those of our previous studies, demonstrating that individual commenters are viewed less positively when outrage goes viral. However, high-status offenders did not elicit less sympathy or generate more positive impressions of those who criticized them, despite the fact that high-status offenders were perceived as committing more harm. People express less favorable impressions of commenters in viral outrage regardless of whether outrage is directed toward a layperson or celebrity.

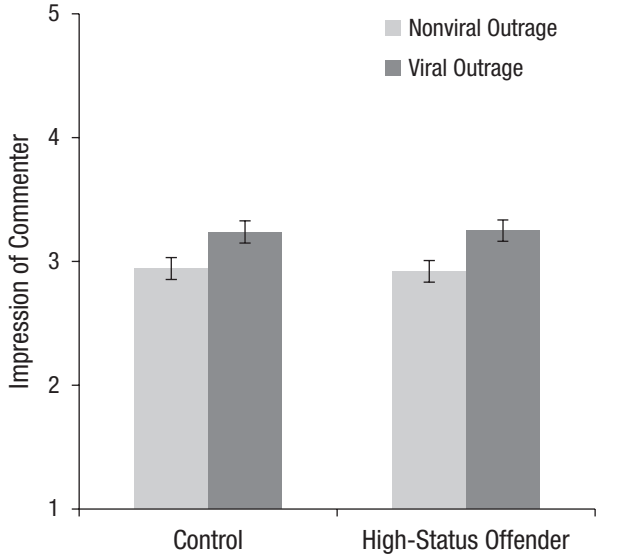

Fig. 9. Results from Study 4a: mean impression of target commenter as a function of offender type and outrage condition. Impression ratings ranged from 1 to 7; higher scores indicate more negative impressions. Error bars represent standard errors.

Study 4b

In Study 4b, we sought to influence sympathy toward the offender using a different manipulation. We theorized that participants would feel less sympathetic toward an offender who was affiliated with a group that most people abhor. Inspired by the 2017 "Unite the Right" rally led by White supremacists and the outrage that followed, we manipulated whether the offender was part of a White supremacist organization. We predicted that viral outrage would fail to inspire more sympathy toward a White supremacist and thus would not lead to more negative impressions of individuals who participated in the outrage.

Method

Participants. Participants ($N = 799$; 443 women; 178 racial-minority individuals; mean age = 37.03 years, $SD = 12.21$) were recruited and compensated through Mechanical Turk. Participants were excluded if their IP address was recorded more than once ($n = 30$), leaving a final sample of 769. As with Study 4a, in Study 4b we set a larger target sample size of 200 per cell because this study was conducted to address questions that arose during the peer-review process. This study was preregistered at https://osf.io/kges3/.

Fig. 10. The offensive stimulus used in Study 4b.

Procedure. Participants saw a tweet taken from a real story (Fig. 10) in which a man belittled an immigrant (Bruk, 2016). We did not use the individual's real name. We manipulated the status of the perpetrator by describing him as either "a college student" (control condition) or "a college student who is a member of a White supremacist organization at his university called the 'White Student Union'" (unsympathetic-offender condition). Participants then received the minimal manipulation of viral (vs. nonviral) outrage in which they read 10 or 2 tweets, respectively, condemning this individual (the same outrage stimuli used in Study 1). Participants reported the extent to which they felt sympathetic toward the offender, as in Studies 1 to 4a (r = .92), and evaluated a randomly selected commenter, as in Studies 2 to 4a (α = .84). To examine why the unsympathetic offender may elicit less sympathy, we included the measure of perceived harm used in Study 4a (α = .53).

Results

Participants found the original transgression to be moderately offensive (M = 4.85, SD = 1.95; 1, *not at all*, to 7, *very*). This did not differ by condition, ps ≥ .106. First, we tested whether the unsympathetic (vs. control) offender would be perceived as committing more harm. We conducted a 2 (outrage: viral or nonviral) × 2 (offender: unsympathetic or control) ANOVA on perceived harm. This revealed a main effect of offender, F(1, 765) = 16.29, p < .001, η_p^2 = .02; participants believed the transgression was more harmful when the offender was unsympathetic (M = 4.36, SD = 1.14) than when the offender was merely a student (M = 4.03, SD = 1.12). There were no effects of outrage, ps ≥ .555.

Next, we conducted a 2 (outrage: viral or nonviral) × 2 (offender: unsympathetic or control) ANOVA on sympathy toward the offender. This revealed a main effect of outrage, F(1, 765) = 25.92, p < .001, η_p^2 = .03; participants felt more sympathetic when outrage was viral (M = 2.13, SD = 1.62) rather than nonviral (M = 1.62, SD = 1.20). There was also a main effect of offender, F(1, 765) = 34.11, p < .001, η_p^2 = .04; participants were less sympathetic toward the unsympathetic offender (M = 1.59, SD = 1.22) than to the student in the control condition (M = 2.18, SD = 1.61). There was no interaction, p = .115.

We then conducted a 2 (outrage: viral or nonviral) × 2 (offender: unsympathetic or control) ANOVA on impressions of the target commenter, with target stimulus as a fixed factor. This revealed a main effect of outrage, F(1, 756) = 3.99, p = .046, η_p^2 = .01 (Fig. 11); participants formed more negative impressions of the target commenter when outrage was viral (M = 3.22, SD = 1.47) rather than nonviral (M = 3.04, SD = 1.40). There was also a main effect of offender, F(1, 756) = 12.32, p < .001, η_p^2 = .02; participants formed less negative impressions of target commenters who expressed outrage toward the unsympathetic offender (M = 2.96, SD = 1.39) than toward the student in the control condition (M = 3.32, SD = 1.47). There was no interaction, p = .720. The effect of viral (vs. nonviral) outrage on impressions of the target commenter were driven by sympathy toward the offender, 95% CI for the indirect effect = [0.158, 0.365], path a: b = 0.51, path b: b = 0.50.

Given that the unsympathetic (vs. control) offender was perceived as committing more harm, we tested whether sympathy continued to mediate our effects even when controlling for perceived harm (sympathy and perceived harm were correlated, r = .21). Consistent with predictions, our analysis showed that sympathy mediated the effect of viral outrage on negative impressions of the

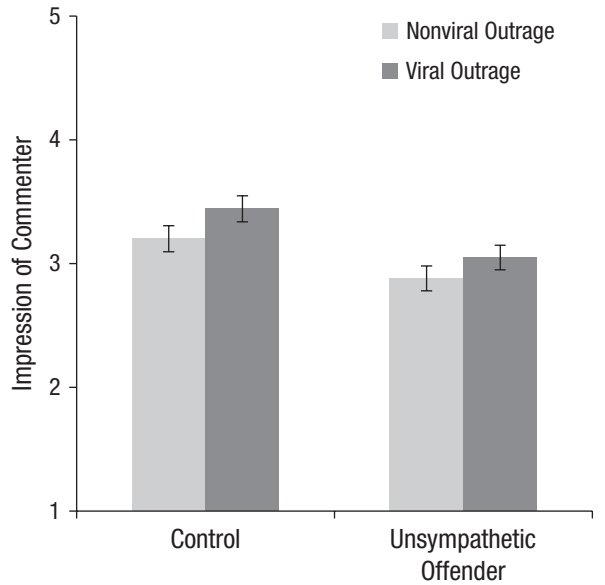

Fig. 11. Results from Study 4b: mean impression of target commenter as a function of offender type and outrage condition. Impression ratings ranged from 1 to 7; higher scores indicate more negative impressions. Error bars represent standard errors.

target commenter even when including perceived harm as a covariate, 95% CI for the indirect effect = [0.136, 0.324], path *a*: $b = 0.51$, path *b*: $b = 0.44$.

We again found that viral outrage led to more negative impressions of individual commenters. Strikingly, although participants were less sympathetic overall toward a White supremacist offender, we did not find our predicted interaction that the effects of viral outrage would be attenuated when the offender was largely unsympathetic. These results suggest that our findings are even more broadly applicable than we had originally anticipated, with viral outrage leading to more negative impressions of individual commenters even when the outrage is directed toward someone as widely despised as a White supremacist.

Study 5

Although outside observers come to regard participants of viral outrage in a more negative light, Ronson (2015) remarked that participants themselves often feel justified in their outrage. In one case, this led to miscommunication and tension when such a participant insisted on her own righteousness while Ronson (2015) attempted to convey why others thought the outrage went too far. Inspired by this, we conducted a final study in which we asked how people who contribute to viral outrage perceive themselves. We hypothesized that individuals likely regard their own expressions of outrage as equally justified regardless of the number of comments that follow and that this is driven by participants feeling less sympathetic for the offender when they themselves have contributed to the outrage. We tested this in Study 5, which contrasted self-perception with perception by observers of the expressed outrage.

Method

Participants. Participants in Phase 1 ($n = 189$; 92 women; 51 racial-minority individuals; mean age = 34.25 years, $SD = 11.84$) and Phase 2 ($n = 190$; 102 women; 53 racial-minority individuals; mean age = 34.05 years, $SD = 11.50$) were recruited and compensated through Mechanical Turk. Participants were excluded if their IP address was recorded more than once ($n = 2$ in Phase 1; $n = 11$ in Phase 2).

Procedure. In Phase 1, participants served as first-person commenters: They saw the offensive tweet used in Study 1 and composed their own tweet in response. Because we were interested in the self-perceptions of those who expressed outrage, we decided a priori to exclude participants who did not write responses condemning the offensive post. We first asked how offensive they found the original tweet (1, *not at all*, to 7, *very*) and excluded participants who responded at or below the midpoint ($n = 66$), as these participants generally wrote messages that did not condemn the offender (i.e., 43 of the 66 participants wrote messages such as "That's pretty funny"). We further excluded participants who responded above the midpoint but wrote messages that did not, on examination, condemn the offender ($n = 20$), leaving a final sample of 101. After writing their message, participants were instructed to imagine that either 2 or 10 additional people had also responded, and they then saw these messages (the responses used in Study 1). They then indicated how sympathetic they felt toward the offender, using the same items as in Studies 2 to 4 ($r = .88$). Finally, they reported the extent to which they felt they themselves were "in the wrong" and "a bully" ($r = .41$).

In Phase 2, another sample of participants served as third-person observers: They first saw the same offensive tweet and read one of the outrage tweets written by a participant in Phase 1 (randomly selected with replacement from the 101 participants who were retained for data analysis in Phase 1). They were then informed that there were 2 or 10 additional commenters and read those messages (again the responses used in Study 1). Next, they indicated how sympathetic they felt for the offender ($r = .81$). Finally, they revisited the first response they saw and rated the extent to which they felt this commenter was "in the wrong" and "a bully" ($r = .87$). As in Phase 1, participants in Phase 2 were excluded if they scored at or below the midpoint of the scale evaluating the offensiveness of the original tweet ($n = 38$), in order to help ensure consistency across samples. This left a final sample of 141.

Results

Participants found the original transgression to be highly offensive ($M = 6.44$, $SD = 0.75$; 1, *not at all*, to 7, *very*). Unexpectedly, participants found the transgression less offensive in the viral-outrage condition ($M = 6.33$, $SD = 0.80$) than in the nonviral-outrage condition ($M = 6.53$, $SD = 0.69$), $F(1, 238) = 6.05$, $p = .015$, $\eta_p^2 = .03$. Because we measured offensiveness prior to any experimental manipulation, this difference by condition indicates a failure of randomization. We thus controlled for perceived offensiveness across our analyses. Our findings remained largely the same regardless of whether we did so, with the exception that the Outrage × Role interaction on sympathy was only marginally significant when we did not include offensiveness as a covariate.

A 2 (outrage: viral or nonviral) × 2 (role: first person or third person) ANCOVA on sympathy toward the offender revealed a main effect of role, $F(1, 237) = 4.69$,

$p = .031$, $\eta_p^2 = .02$; participants were less sympathetic toward the offender when they were first-person commenters ($M = 1.76$, $SD = 1.34$) rather than third-person observers ($M = 2.15$, $SD = 1.48$). There was also a main effect of outrage, $F(1, 237) = 7.45$, $p = .007$, $\eta_p^2 = .03$; participants were more sympathetic when outrage was viral ($M = 2.36$, $SD = 1.62$) rather than nonviral ($M = 1.69$, $SD = 1.20$). Finally, there was a significant interaction, $F(1, 237) = 6.96$, $p = .009$, $\eta_p^2 = .03$. Although third-person observers were more sympathetic toward the offender when outrage was viral ($M = 2.68$, $SD = 1.69$) rather than nonviral ($M = 1.74$, $SD = 1.16$), $p < .001$, first-person commenters were not affected by whether outrage was viral ($M = 1.92$, $SD = 1.43$) or nonviral ($M = 1.62$, $SD = 1.27$), $p = .952$.

Next, a 2 (outrage: viral or nonviral) × 2 (role: first person or third person) ANCOVA on impressions of the target commenter revealed a main effect of role, $F(1, 237) = 29.05$, $p < .001$, $\eta_p^2 = .11$ (Fig. 12): Participants reported less negative impressions of the target commenter when they were in the first-person role ($M = 1.35$, $SD = 0.83$) than when they were in the third-person role ($M = 2.38$, $SD = 1.86$). Importantly, this was qualified by a significant interaction, $F(1, 237) = 5.49$, $p = .020$, $\eta_p^2 = .02$. As in previous studies, third-person observers reported more negative impressions of the target commenter when outrage was viral ($M = 2.76$, $SD = 2.03$) rather than nonviral ($M = 2.08$, $SD = 1.67$), $p = .008$. By contrast, first-person commenters were not affected by whether outrage was viral ($M = 1.25$, $SD = 0.56$) or nonviral ($M = 1.43$, $SD = 1.00$), $p = .413$. These effects were mediated by sympathy toward the offender, 95% CI for the indirect effect = [0.074, 0.398], path *a*: $b = 0.55$, path *b*: $b = 0.36$. As more people expressed outrage, first-person commenters continued to view themselves in a positive light, even as third-person observers' perceptions of them soured.

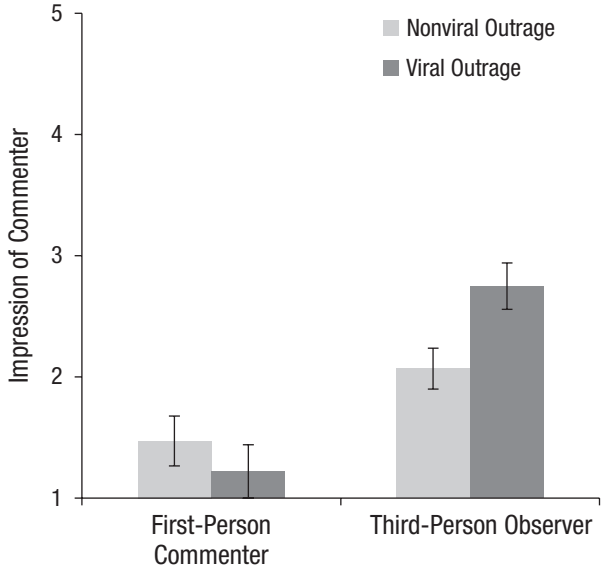

Fig. 12. Results from Study 5: mean impression of target commenter as a function of participants' role and outrage condition. Impression ratings ranged from 1 to 7; higher scores indicate more negative impressions. Error bars represent standard errors.

General Discussion

Even when moral outrage toward racist, sexist, or otherwise insensitive remarks seems appropriate and justified in isolation, observers form more negative impressions of individual commenters when the outrage goes viral (Table 1). This is because viral outrage elicits more sympathy toward the offender, tainting perceptions of individuals who contribute to this aggregated punishment. Furthermore, viral outrage creates a widening gap between how commenters see themselves and how they are seen by others, as individuals who participate in viral outrage continue to believe they are in the right, even as outside observers come to disagree.

Referring to online shaming, one author wrote, "The snowflake never needs to feel responsible for the avalanche" (Ronson, 2015, p. 284). But we have shown that every snowflake is held accountable for an avalanche. Expressions of public outrage are an important component of social life (Tannenbaum, Uhlmann, & Diermeier, 2011), serving throughout history to clarify and reassert group norms and boundaries (Erikson, 1966). Yet viral outrage leads to the paradox that the same individual outrage that seems laudable and necessary in isolation may be viewed more as excessive and bullying when echoed by multitudinous other users.

This paradox is not limited to the digital domain. Whenever individuals inflict bilateral punishment on a group member, they risk contributing to excessive aggregate punishment. If a child is excluded from every single birthday party for biting a classmate, each parent may be justified in withholding his or her invitation, but the result is complete ostracism, a punishment seemingly out of proportion with the transgression. Similarly, if someone no longer asks a coworker to lunch for making an offensive remark at the company retreat, he or she may be and feel perfectly justified, but when everyone on the floor makes the same individual decision, each of them now seems more cruel for doing so. Such aggregate punishment is also evident in students' moral outrage on American college campuses (Hartocollis, 2015).

How can a person condemn injustice without suffering backlash? Our findings do not provide easy solutions. In Study 4a, commenters who participated in viral

outrage toward a high-status public figure (rather than a layperson) were just as likely to be viewed with suspicion. In Study 4b, although outrage toward a White supremacist was considered more praiseworthy overall, viral outrage toward such an offender nonetheless increased negative impressions of individual commenters. Future research should identify potential boundary conditions for our findings, as the present studies failed to identify moderators to attenuate the effects of viral outrage.

The question of how best to respond to injustice in the digital age remains largely open. The exponential nature of online moral outrage may significantly alter how outrage is elicited, expressed, and perceived (Crockett, 2017), raising the need for more research to clarify the social costs and ramifications of this novel phenomenon. The challenge lies in reconciling the counterintuitive notion that a collection of individually praiseworthy actions may cumulatively result in an unjust outcome.

Action Editor

Ayse K. Uskul served as action editor for this article.

Author Contributions

T. Sawaoka developed the study concept. Both authors contributed to the study design. Testing, data collection, and data analysis were performed by T. Sawaoka. T. Sawaoka drafted the manuscript, and B. Monin provided critical revisions. Both authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618780658

Open Practices

All data and materials have been made publicly available via the Open Science Framework (OSF) and can be accessed at https://osf.io/qtsx2/. The design and analysis plans for Studies 4a and 4b were preregistered at the OSF and can be accessed at https://osf.io/kges3/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618780658. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368–378.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution & Human Behavior*, *27*, 325–344.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*, 107–122.

Blistein, J. (2015, March 31). 'Daily Show' hire Trevor Noah under fire for tweets offending Jews, women. *Rolling Stone*. Retrieved from http://www.rollingstone.com/tv/news/trevor-noah-under-fire-for-tweets-offending-jews-women-20150331

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, USA*, *114*, 7313–7318.

Bruk, D. (2016, August 18). Blake Shelton apologizes for offensive tweets, sort of. *Cosmopolitan*. Retrieved from http://www.cosmopolitan.com/entertainment/news/a62943/blake-shelton-racist-tweets-apology/

Cialdini, R. B. (1993). *Influence: Science and practice* (3rd ed.). New York, NY: HarperCollins.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026.

Clemons, T. (2016, September 30). 'Blackface' social media post goes viral, lands Prairie View A&M athlete in hot water. *ABC News*. Retrieved from http://abc13.com/news/prairie-view-athlete-in-hot-water-after-blackface-post/1534688/

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behavior*, *1*, 769–771.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*, 784–803.

Erikson, K. T. (1966). *Wayward puritans: A study in the sociology of deviance*. London, England: John Wiley & Sons.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Fragale, A. R., Rosen, B., Xu, C., & Merideth, I. (2009). The higher they are, the harder they fall: The effects of wrongdoer status on observer punishment recommendations and intentionality attributions. *Organizational Behavior and Human Decision Processes*, *108*, 53–65.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford, England: Oxford University Press.

Hartocollis, A. (2015, December 7). Yale lecturer resigns after email on Halloween costumes. *The New York Times*. Retrieved

from https://www.nytimes.com/2015/12/08/us/yale-lecturer-resigns-after-email-on-halloween-costumes.html

Hechler, S., & Kessler, T. (2018). On the difference between moral outrage and empathic anger: Anger about wrongful deeds or harmful consequences. *Journal of Experimental Social Psychology*, *76*, 270–282.

Hempel, J. (2017, October 18). The problem with #metoo and viral outrage. *Wired*. Retrieved from https://www.wired.com/story/the-problem-with-me-too-and-viral-outrage/

Inesi, M. E., Gruenfeld, D. H., & Galinsky, A. D. (2012). How power corrupts relationships: Cynical attributions for others' generous acts. *Journal of Experimental Social Psychology*, *48*, 795–803.

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, *95*, 76–93.

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, *138*, 535–545.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164.

Quinn, B. (2015, June 11). Nobel laureate Tim Hunt resigns after 'trouble with girls' comment. *The Guardian*. Retrieved from https://www.theguardian.com/education/2015/jun/11/nobel-laureate-sir-tim-hunt-resigns-trouble-with-girls-comments

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, *30*, 98–103.

Rattan, A., & Dweck, C. S. (2010). Who confronts prejudice? The role of implicit theories in the motivation to confront prejudice. *Psychological Science*, *21*, 952–959.

Ronson, J. (2015). *So you've been publicly shamed*. New York, NY: Penguin.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*, 574–586.

Salerno, J. M., & Peter-Hagene, L. C. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science*, *24*, 2069–2078.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January). *Life after* p-*hacking*. Paper presented at the Meeting of the Society for Personality and Social Psychology, New Orleans, LA.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845–851.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*, 1249–1254.

Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy & Public Affairs*, *44*, 193–217.

Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: Just, brave, and caring. *Journal of Personality and Social Psychology*, *86*, 629–647.